KTH Computer Science
and Communication

Exam in DD2421 Machine Learning
2017-10-21, kl 9.00 – 13.00

Aids allowed: *calculator*, *language dictionary*.

In order to pass, you have to fulfill the requirements defined both in the A- and in the B-section.

A Questions on essential concepts

Note: As a prerequisite for passing you must choose the correct answer to almost *all* questions. Only *one* error will be accepted, so pay good attention here.

A-1 Probabilistic Learning

The goal of *maximum a posteriori* estimation is to find the model parameters that ...

a) optimize the likelihood of the new observations in conjunction with the a priori information.

b) maximize a convex optimality criterion.

c) maximize the prior.

A-2 Naive Bayes Classifier

What is the underlying assumption unique to a *naive Bayes* classifier?

a) All features are regarded as conditionally independent.

b) A Gaussian distribution is assumed for the feature values.

c) The number of features (the dimension of feature space) is large.

A-3 Shannon Entropy

Consider a single toss of fair coin. Regarding the uncertainty of the outcome {head, tail}, the entropy is equal to ...

a) zero bit.

b) one bit.

c) two bits.

A-4 Regression

In regression, *regularization* can be achieved by adding a term, so-called *shrinkage penalty*. Which one of the methods below introduces the additional term.

a) Least squares.

b) Ridge regression.

c) k-NN regression.

A-5 Perceptron Learning Rule

The *Perceptron Learning Rule* is used to ...

a) adjust the step size for optimal learning.

b) update the weights when a training sample is erroneously classified.

c) minimize the entropy over the whole training dataset.

A-6 Support Vector Machine

What property of the *Support Vector Machine* makes it possible to use the *Kernel Trick*?

a) The weights are non-zero only in a limited part of the state space.

b) The margin width grows linearly with the number of sample points.

c) The only operation needed in the high dimensional space is to compute scalar products between pairs of samples.

A-7 Ensemble Learning

Which one below correctly describes the property of *Adaboost Algorithm* for classification?

a) Adaboost algorithm is more suited to multi-class classification than binary classification.

b) Models to be combined are requied to be as similar as possible to each other.

c) A weight is given to each training sample, and it is iteratively updated.

A-8 Principal Component Analysis (PCA)

All of the following statements about PCA are true *except*

a) PCA serves for subspace methods to represent the data distribution in each class.

b) PCA is useful for reducing the effective dimensionality of data.

c) PCA is a supervised learning method that requires labeled data.

Note: Your answers (eight of them) need be on a solution sheet (**this page will not be received**).

B Graded problems

A pass is guaranteed with the required points for 'E' below in this section *and* the prerequisite in the A-section.

Preliminary number of points required for different grades:

$$
\begin{aligned}
24 \le p \le 27 &\rightarrow A \\
20 \le p < 24 &\rightarrow B \\
16 \le p < 20 &\rightarrow C \\
12 \le p < 16 &\rightarrow D \\
9 \le p < 12 &\rightarrow E \\
0 \le p < 9 &\rightarrow F
\end{aligned}
$$

B-1 Terminology (4p)

For each term (a–h) in the left list, find the explanation from the right list which *best* describes how the term is used in machine learning.

a) Error backpropagation

b) Expectation Maximization

c) k-fold cross validation

d) The Lasso

e) k-means

f) RANSAC

g) Subspace

h) Fisher's criterion

1) An approach to find useful dimension for classification

2) Algorithm to learn with latent variables

3) A space spanned by a set of linearly independent vectors

4) Estimating expected value

5) An approach to train artificial neural networks

6) Random strategy for amplitude compensation

7) A strategey to generate k different models

8) The last solution

9) Method for estimating the mean of k observations

10) Algorithm to estimate errors

11) Robust method to fit a model to data with outliers

12) An approach to regression that results in feature seletion

13) Clustering method based on centroids

14) A subportion of area defined by two sets of parallel lines

15) A technique for assessing a model while exploiting available data for training and testing

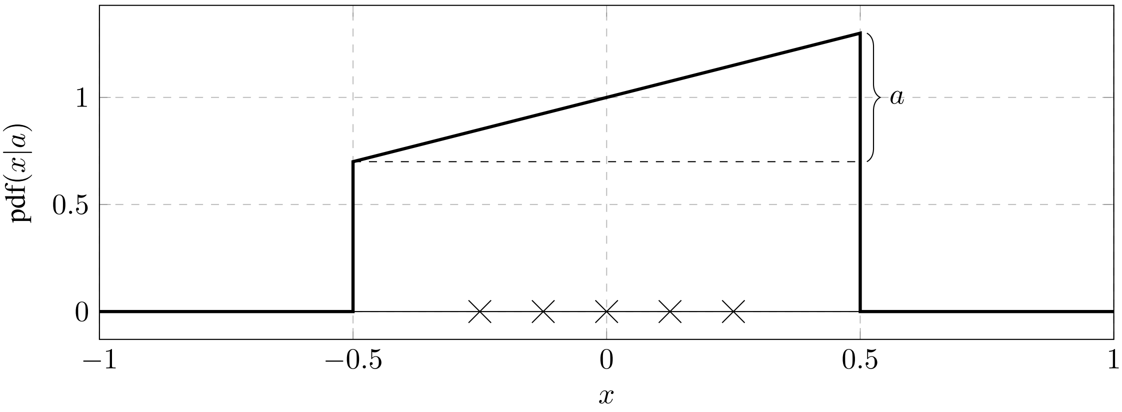


Figure 1. Illustration for Problem B-2.

B-2 Probability based learning (3p)

The continuous probability distribution function (PDF) depicted in Figure 2 depends on one parameter a related to the slope of the line and can be defined as:

$$\text{pdf}(x|a) = \begin{cases} 1 + ax, & \text{for } -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

The figure also shows five data points with x-coordinates $-\frac{1}{4}, -\frac{1}{8}, 0, \frac{1}{8},$ and, $\frac{1}{4}$ that are considered to be independently drawn from the distribution. We call this set of points $\mathcal{D}$.

a) What is the range of values for a to ensure that the above definition is a valid probability distribution function?

b) Using the likelihood of the data $\mathcal{D}$ given the model parameter a, select the model that best fits the data between the following three alternatives: $a = 0$, $a = 1$, and $a = -2$. (If you do not have a calculator, use fractions.)

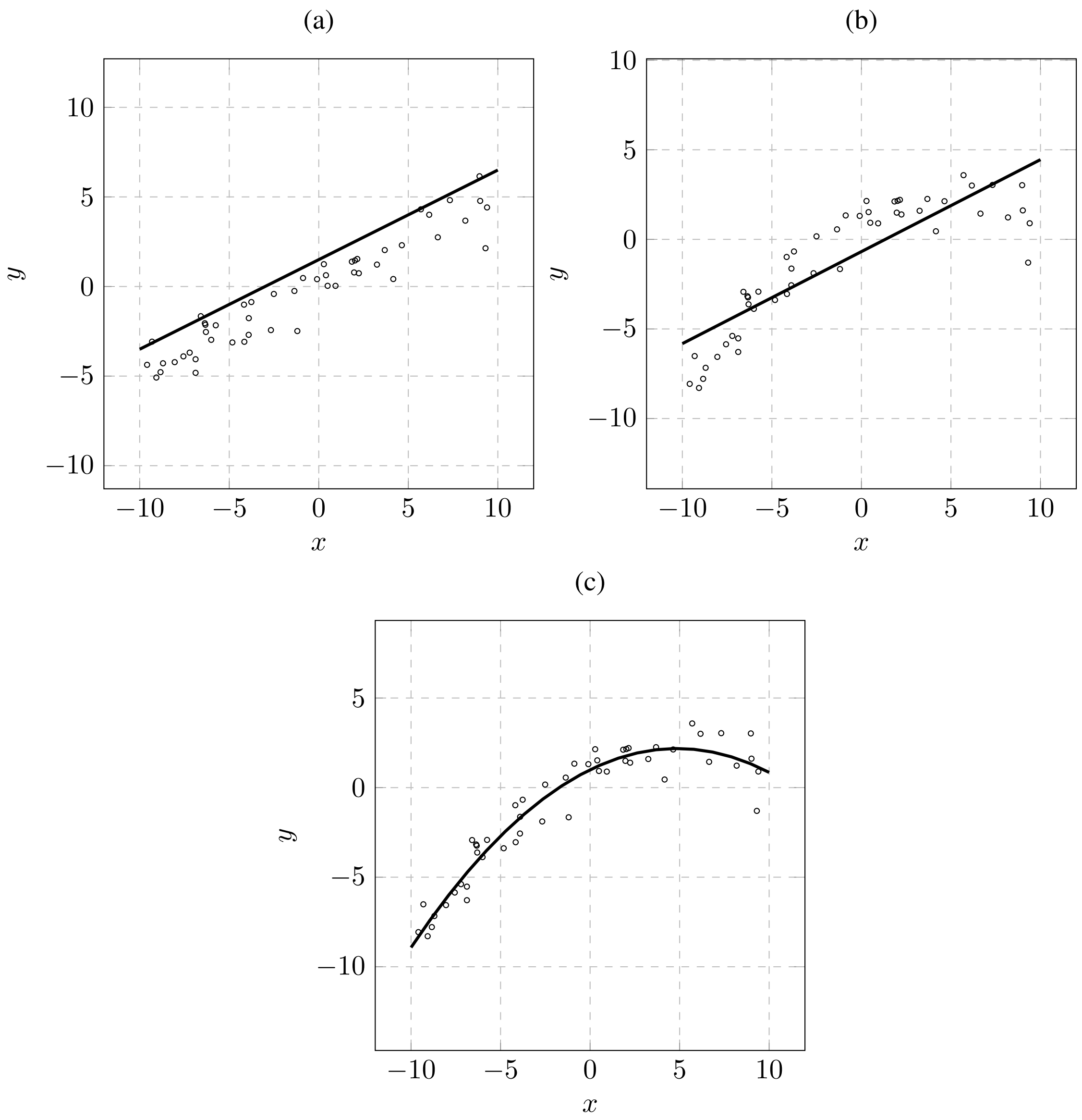
c) Is the best model you found at the previous point also the best over all possible values of a? Motivate your answer.

B-3 Probability based Learning (3p)

For each of the following cases, determine if the illustration can correspond to a case of probabilistic linear regression with:

- error (residual) distributed according to $\mathcal{N}(0, \sigma^2)$, and
- model parameters obtained by maximum likelihood estimation using the data points in the illustration.

Motivate your answer for each case (answers without motivation receive zero points).



B-4 Classification (3p)

Suppose that we take a data set, divide it into training and test sets, and then try out two different classification procedures. We use *two-thirds* of the data for training, and the remaining *one-third* for testing. First we use Logistic Regression and get an error rate of 10% on the training data. We also get the average error rate (averaged over both test and training data sets) of 15%. Next we use k-nearest neighbor (where $k = 1$) and get an average error rate (averaged over both test and training data sets) of 10%.

a) What was the error rate with 1-nearest neighbor on the test set?

b) What was the error rate with the Logistic Regression on the test set?

c) Based on these results, indicate the method which we should prefer to use for classification of new observations, with a simple reasoning.

B-5 Random Forests (2p)

Choose the correct answers in the following questions on Random Forests.

a) Mainly two kinds of randomness are known to form the basic principle of Random Forests. In which two of the following processes are those randomnesses involved?

i. In generating bootstrap replicas.
ii. In deciding the number of trees used.
iii. In feature selection at each node.
iv. In the way to formulate the information gain.
v. In the rule of terminating a node as a leaf node.
vi. In combining the results from multiple trees.

Simply indicate two among those above.

b) Suppose we have generated a Random Forest using five bootstrapped samples from a data set containing three classes, {Green, Blue, Red}. We then applied the forest to a specific test input, x, and observed five estimates of $P(\text{Class is Blue}|x)$: 0.4, 0.4, 0.6, 0.65, and 0.7.

Consider two common ways to combine these results together into a single class prediction: the majority vote approach, and the other based on the average probability. In this example, what is the final classification under each of these two approaches?

i. Green or Red in both approaches.
ii. Green or Red in averaging and Blue in majority vote.
iii. Blue in both approaches.

Indicate one among the above, and motivate your answer by short phrases.

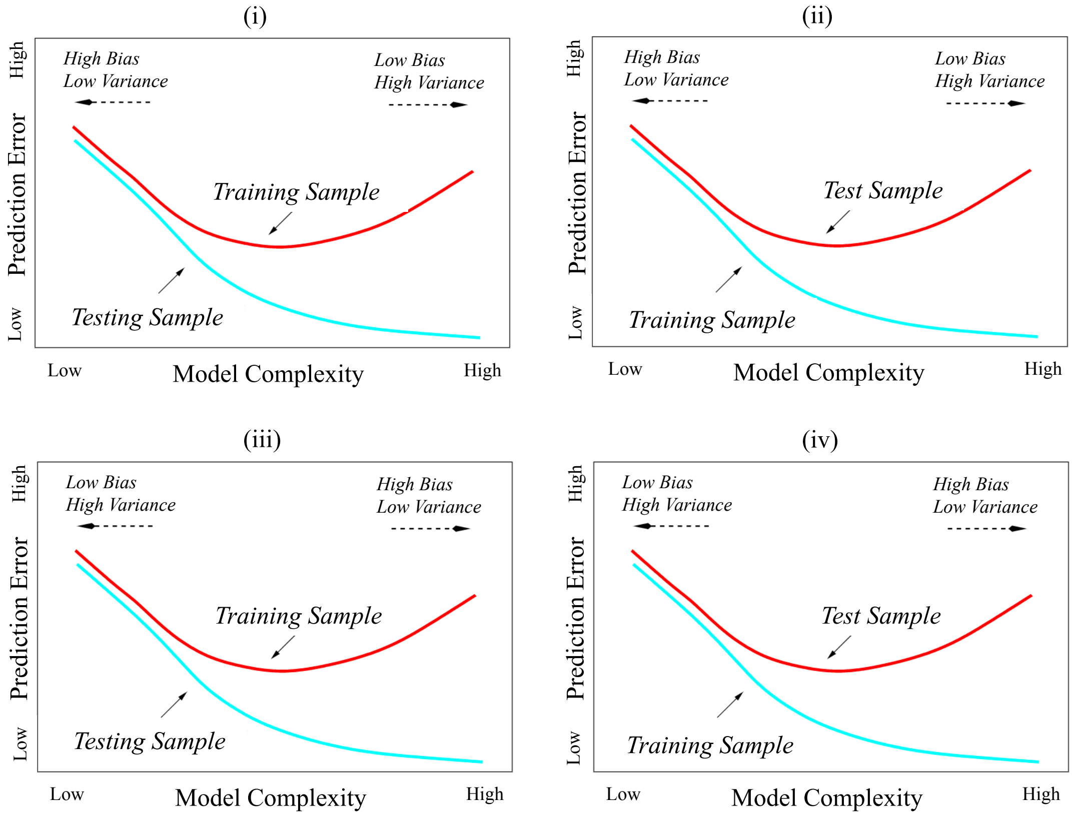


Figure 2. Graphs for Problem B-6.

B-6 Bias and Variance (3p)

a) One of the four subfigures (i)-(iv) in Figure 2 displays the typical trend of prediction error of a model for training and testing data with comments on its bias and variance, {high, low}. Which one of the four figures most well represents the general situation?

b) Now consider the specific case of using *Bagging* by an ensemble of decision tree classifiers. What sort of improvememt can be expected in the ensemble predictions in terms of *bias* or *variance* of the classifier as a whole?

c) Briefly explain the main reason why the prediction errors have different trend for training samples and test samples.

B-7 Support Vector Machines (3p)

Training a support vector machine using a quadratic kernel

$$\mathcal{K}(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + 1)^2$$

has resulted in the following four support vectors:

$$s_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad s_2 = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} \qquad s_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad s_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

The first two (s_1 and s_2) are positive samples while the other two (s_3 and s_4) are negative samples. The corresponding α-values are: $\alpha_1 = \alpha_2 = \frac{7}{16} = 0.4375$ and $\alpha_3 = \alpha_3 = \frac{3}{8} = 0.375$.

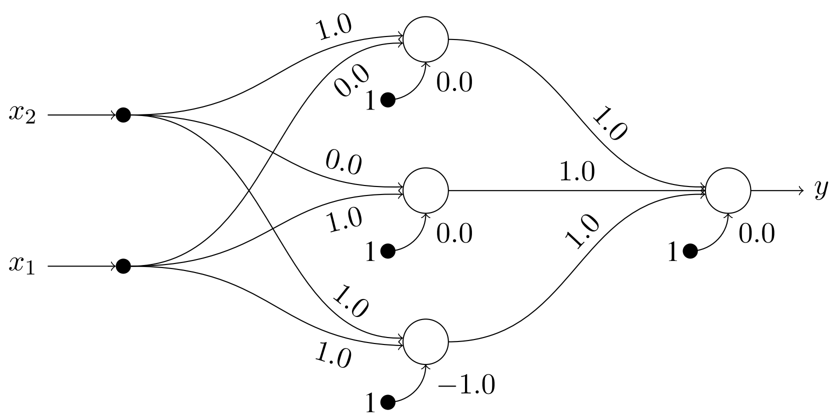
Determine how the following *new* datapoints will be classified:

$$x_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

You must show the formulas and calculations used to arrive at your answer.

B-8 Artificial Neural Networks (3p)

Consider a feed-forward neural network with threshold units. The number of nodes and all weight values are shown in the figure. Circles indicate threshold units (with threshold at zero and output $\in \{-1, 1\}$) while the small filled circles are just "pass through" nodes.



a) Draw a diagram of the input space and show the position of the separating hyperplanes implemented by the *hidden units*.

b) In the same figure, indicate the area where the output will be high ($y = 1$) by shading it.

B-9 Curse of Dimensionality (3p)

Answer the following questions regarding the phenomenon known as the curse of dimentionality; when the number of features p is large, there tends to be a deterioration in the performance of some approaches such as k-nearest neighbours.

Suppose that we have a set of observations, each with measurements on $p = 1$ feature, x. We assume that x is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of x closest to that test observation. For instance, in order to predict the response for a test observation with $x = 0.6$, we will use observations in the range [0.55, 0.65]. On average, the fraction of the available observations we will use to make the prediction can be considered as 10%, ignoring the range $x < 0.05$ and $x > 0.95$.

a) Suppose that we have a set of observations, each with measurements on $p = 2$ features, x_1 and x_2. We assume that (x_1, x_2) are uniformly distributed on [0, 1] $\times$ [0, 1]. We wish to predict a test observation's response using only observations that are within 10% of the range of x_1 and within 10% of the range of x_2 closest to that test observation.

On average, what fraction of the available observations will we use to make the prediction?

b) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value [0, 1]. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation.

What fraction of the available observations will we use to make the prediction?

c) Furthermore, suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For p = D features, what is the length, l, of each side of the hypercube? Comment on your answer.